CLAIMS

We claim:

1. A computer system having one or more memories and one or more central processing units (CPUs), the system comprising:

one or more multimedia items, stored in the memories, each multimedia item having two or more disparate modalities, the disparate modalities being at least one or more visual modalities and one or more textual modalities; and

a combining process that creates a visual feature vector for each of the visual modalities and a textual feature vector for each of the textual modalities, and concatenates the visual feature vectors and the textual feature vectors into a unified feature vector.

2. A system, as in claim 1, further comprising a classifier induction process that induces a classifier from the unified feature vectors.

3. A system, as in claim 2, where the classifiers include any one or more of the following: a hyperplane classifier, a rule-based classifier, a Bayesian classifier, maximum likelihood classifier.

4. A system, as in claim 1, further comprises:

one or more classifiers having one or more classes; and

an application process that for each of the multimedia items, uses the classifiers to predict zero or more of the classes to which the respective multimedia items belong, the multimedia items being unprocessed multimedia items, and where in the case that zero categories are predicted the multimedia item does not belong to any class.

5. A system, as in claim 1, further comprises a transformation process that transforms one or more feature vectors in the set of visual feature vectors and textual feature vectors in order to make one or of more the visual feature vectors compatible with one or more of the textual feature vectors for the all multimedia items.

6. A system, as in claim 5, where the visual feature vectors and textual feature vectors are made compatible by limiting the component values in the respective visual and textual feature vectors.

7. A system, as in claim 6, where the component values includes: a binary value; a one bit binary value; a 0, 1, 2 or many value; a value in a range of discrete value; and a 0, 1, 2, or 3 value.

8. A system, as in claim 5, where the visual feature vectors and textual feature vectors are made compatible by limiting the difference between the magnitude of the visual and textual feature vectors.

9. A system, as in claim 8, where the difference in magnitude is limited by normalizing the visual and textual vectors.

10. A system, as in claim 5, where the visual feature vectors and textual feature vectors are made compatible by limiting the difference between the number of components in the respective vectors.

11. A system, as in claim 1, where the visual feature vectors comprise one or more of the following: a set of ordered components, a set of unordered components, a set of only temporally ordered components, a set of only spatially ordered components, a set of temporally and spatially ordered components, a set of visual features extracted from ordered key intervals, a set of visual features extracted from ordered key intervals divided into regions, and a set of semantic features.

12. A system, as in claim 1, where there visual feature vectors has a fixed length, the fixed length being independent of the length of the multimedia items.

13. A system, as in claim 1, where the visual feature vectors comprise one or more components that are selected so that the visual feature vector is sparse.

14. A system, as in claim 1, where the visual feature vectors represent any one or more of the following: a color, a motion, a visual texture, an optical flow, a semantic meaning, semantic meanings derived from one or more video streams, an edge density, a hue, an amplitude, a frequency, and a brightness.

15. A system, as in claim 1, where the textual feature vectors are derived from any one or more of the following: close captions, open captions, captions, speech recognition applied to one or more audio input, semantic meanings derived from one or more audio streams, and global text information associated with the media item.

16. A computer system having one or more memories and one or more central processing units (CPUs), the system comprising:

one or more multimedia items, stored in the memories, each multimedia item having two or more disparate modalities, the disparate modalities being at least one or more visual modalities and one or more textual modalities;

a block process that divides the multimedia items into blocks of one or more key intervals, each key interval having one more frames of the multimedia items;

a combining process that creates a visual feature vector for each of the visual modalities and a textual feature vector for each of the textual modalities, and concatenates the visual feature vectors and the textual feature vectors into a unified feature vector;

one or more classifiers having one or more classes;

an application process that for each of the blocks, uses the classifiers to determine zero or more of the classes to which the respective blocks belong to; and

a segmentation process that finds temporally contiguous groups of the blocks and combines the contiguous groups into media segments where all the blocks in the media segment have one or more of the same classes.

17. A system, as in claim 16, further comprising an aggregation process that aggregates two or more of the media segments belonging to the same class with one or more media segments of a different class according to one or more aggregation rules.

18. A system, as in claim 17, where the aggregation rules include any one or more of the following rule types: segment region rules, segment boundary indicator rules, and learned rules that are derived from training data.

19. A system, as in claim 18, where the segment region rule has a minimum segment length constraint and a plurality of rules that change small sequences of blocks of varying categorization into blocks of equal category.

20. A system, as in claim 18, where the segment boundary indicator rules are multimedia cues and these multimedia cues are one or more of the following: a shot transition, an audio silence, a speaker change, an end-of-sentence in speech transcript, and a topic change indicator in the closed-caption.

21. A system, as in claim 18, where the learned rules are the costs of transitions and the aggregations process aggregates two or more of the media segments belonging to the same class with one or more media segments of a different class by minimizing the overall cost of the sequence of segments.

22. A method for segmenting multimedia streams comprising the steps of:

storing one or more multimedia items in one or more memories of computer, each multimedia item having two or more disparate modalities, the disparate modalities being at least one or more visual modalities and one or more textual modalities;

dividing the multimedia items into blocks of one or more key intervals, each key interval having one more frames of the multimedia items;

for each block, creating a visual feature vector for each of the visual modalities and a textual feature vector for each of the textual modalities;

for each block, concatenating the visual feature vectors and the textual feature vectors into a unified feature vector;

categorizing each of the blocks by categorizing the respective unified feature vector; and

assembling two or more of the categorized blocks into a segment.

23. A memory storing a program, the program comprising the steps of:

storing one or more multimedia items in one or more memories of computer, each multimedia item having two or more disparate modalities, the disparate modalities being at least one or more visual modalities and one or more textual modalities;

dividing the multimedia items into blocks of one or more key intervals, each key interval having one more frames of the multimedia items;

for each block, creating a visual feature vector for each of the visual modalities and a textual feature vector for each of the textual modalities;

for each block, concatenating the visual feature vectors and the textual feature vectors into a unified feature vector;

categorizing each of the blocks by categorizing the respective unified feature vector; and

assembling two or more of the categorized blocks into a segment.

24. A system for segmenting multimedia streams comprising:

means for storing one or more multimedia items in one or more memories of computer, each multimedia item having two or more disparate modalities, the disparate modalities being at least one or more visual modalities and one or more textual modalities;

means for dividing the multimedia items into blocks of one or more key intervals, each key interval having one more frames of the multimedia items;

means for creating a visual feature vector for each of the visual modalities and a textual feature vector for each of the textual modalities, block by block;

means for concatenating the visual feature vectors and the textual feature vectors into a unified feature vector, block by block;

means for categorizing each of the blocks by categorizing the respective unified feature vector; and

means for assembling two or more of the categorized blocks into a segment.